LIMSI @ 2014 Clinical Decision Support Track

Eva D'hondt[*], Brigitte Grau[*], Stéfan Darmoni[**], Aurélie Névéol[*], Matthieu Schuers[**] and Pierre Zweigenbaum[*]

[*]LIMSI (CNRS UPR 3251) 91405 Orsay, France
[**]CISMeF-TIBS-LITIS EA 4108, Rouen University Hospital 76031 Rouen, France

Abstract

In this paper we present our participation in the 2014 TREC Clinical Decision Support Track. The goal of this track is to find relevant medical literature for a case report which should help address one specific clinical aspect of the case. Since it was the first time we participated in this task, we opted for an exploratory approach to test the impact of retrieval systems based on Bag-of-Words (BoW) or Medical Subject Headings (MeSH) index terms. In all five submitted runs, we used manually constructed MeSH queries to filter a target corpus for each of the three clinical question types. Query expansion (for both MeSH and BoW runs) was based on the automatic generation of disease hypotheses for which we used data from OrphaNet [4] and the Disease Symptom Knowledge Database [3]. Our best run was a MeSH-based run in which PubMed was queried directly with the MeSH terms extracted from the case reports, combined with the MeSH terms of the top 5 disease hypotheses generated for the case reports. Compared to the other participants we achieved low scores. Preliminary analysis shows that our corpus filtering method was too strict and has a negative impact on recall.

Keywords

Document Retrieval, UMLS, MeSH, Clinical Decision Support

1 Introduction

The goal of the Clinical Decision Support Track is to retrieve relevant biomedical articles given a patient record. This year is the first time that this particular track has been organized. In this instalment the patient records are short case reports that describe a medical case, for example,

> A woman in her mid-30s presented with dyspnea and hemoptysis. CT scan revealed a cystic mass in the right lower lobe. Before she received treatment, she developed right arm weakness and aphasia. She was treated, but four years later suffered another stroke. Follow-up CT scan showed multiple new cystic lesions.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
LIMSI @ 2014 Clinical Decision Support Track	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Computer Science Laboratory for Mechanics and Engineering Sciences (LIMSI),(CNRS UPR 3251),91405 Orsay, France,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

presented in the proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014) held in Gaithersburg, Maryland, November 19-21, 2014. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA).

14. ABSTRACT

In this paper we present our participation in the 2014 TREC Clinical Decision Support Track. The goal of this track is to nd relevant medical literature for a case report which should help address one speci c clinical aspect of the case. Since it was the rst time we participated in this task, we opted for an exploratory approach to test the impact of retrieval systems based on Bag-of-Words (BoW) or Medical Subject Headings (MeSH) index terms. In all ve submitted runs, we used manually constructed MeSH queries to lter a target corpus for each of the three clinical question types. Query expansion (for both MeSH and BoW runs) was based on the automatic generation of disease hypotheses for which we used data from OrphaNet [4] and the Disease Symptom Knowledge Database [3]. Our best run was a MeSH-based run in which PubMed was queried directly with the MeSH terms extracted from the case reports, combined with the MeSH terms of the top 5 disease hypotheses generated for the case reports. Compared to the other participants we achieved low scores. Preliminary analysis shows that our corpus ltering method was too strict and has a negative impact on recall.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**7**	

The case reports are available as short summaries or slightly longer case descriptions. In the experiments reported below we only used the case summaries.

The concept of 'relevancy' in this task differs from that in standard document retrieval: The developed system should not just find documents that refer to the illness or symptoms described in the case report, but that also answer a clinical question such as "What is the patient's diagnosis?". In total there were three different clinical question types. Each case report had one associated clinical questions.

1. What is the patient's diagnosis? (diagnosis)
2. How should the patient be treated? (treatment)
3. Which tests should be prescribed to treat the patient? (test)

The target document collection is a 21-01-2014 snapshot from the Open Access Subset from PubMed Central (PMC). It contains 733,138 articles which were furnished in NXML format by the organizers.

A complicating factor in this year's task was the lack of official training material, that is, Gold Standard data on links between case reports and relevant journal articles in the Open Access set, according to the clinical question type. We looked into using material from the CasesDatabase website[1], a website that aggregated case reports from different medical journals and used text mining tools to automatically extract condition, symptom, intervention, pathogen, patient demographic and other data fields from the articles. With this information the case reports could be grouped per particular symptom or condition. While initially promising, in practice the CasesDatabase proved difficult to use as a training set since it is not representative for our dataset: The articles contained in the CasesDatabase were all case reports, while the Open Access set is much more diverse. Moreover the case descriptions in the case reports were always not representative for the shorter description that make up the topics in the Clinical Decision Track. Due to time constraints we were not able to add extra information and turn it into a usable training set. The systems that were created for this task are therefore largely untrained.

The use of MeSH terms as index terms sets PubMed apart from other document collections. The majority of articles in PubMed are manually indexed by specialists with 5 to 15 MeSH terms that are considered the most pertinent and detailed for the subjects discussed. For example, an article describing a treatment of the Kawasaki Disease will generally be indexed with the MeSH heading for this disease "mucocutaneous lymph node syndrome" and the MeSH subheading "drug therapy". However, the specialists will often not add MeSH terms for the symptoms associated with the disease since many symptoms are not linked to one particular disease and therefore hold little classification value[2]. In our participation in the Clinical Decision Support track we were interested in the impact of using MeSH terms, which are more precise, versus a Bag of Words approach when expanding the query with disease hypotheses. These hypotheses were generated by a self-constructed Diagnostic Clinical Decision Support system, hereafter referred to as the Symptom Checker. The Symptom Checker takes a set of symptoms (extracted from the case report) as input, and returns a ranked list of disease hypotheses.

This paper is organised as follows. In Section 2, we describe the individual components of the different systems that were built in the course of this track. In Section 3, we present the five runs

[1] Could be found at `http://www.casesdatabase.com/` but has gone offline since July 2014.

[2] A noted example are case reports which may be indexed for symptoms since they often describe atypical occurrences of a disease or condition.

that were submitted for evaluation. Results are presented in Section 4, and the conclusions are discussed in Section 5.

2 System Components

2.1 Clinical question types as MeSH queries

The clinical question aspect poses an interesting problem for retrieval: A document is only relevant if it refers to the disease or condition(s) described in the topic query *and* if it contains enough information that is useful for answering the clinical question.

We hypothesized that documents which are indexed with MeSH terms that pertain to the clinical questions, e.g. "diagnosis"[Subheading] for the diagnosis question, would be the most important to find while other documents that may refer to the same disease but not contain information specific to the clinical question would only constitute noise, especially in the Bag-of-Word experiments (see infra).

We therefore opted for a filtering approach in which we translated the clinical question types into MeSH queries that were used to query PubMed Open Access subset online. The returned result sets were then filtered, only keeping those documents that also appeared in the provided snapshot. This process resulted in three different subcorpora (hereafter referred to as 'diagnosis corpus', 'test corpus' and 'treatment corpus') that were used for further experiments.

We used the following MeSH queries:

- Diagnosis:

  "diagnosis"[MeSH Terms] OR "diagnosis, oral"[MeSH Terms] OR "diagnostic equipment"[MeSH Terms] OR "diagnostic services"[MeSH Terms] OR "nursing diagnosis"[MeSH Terms] OR "reagent kits, diagnostic"[MeSH Terms] OR "diagnosis"[Subheading] OR "diagnostic use"[Subheading]

- Treatment

  "psychiatric somatic therapies"[MeSH Terms] OR "psychotherapy"[MeSH Terms] OR "root canal therapy"[MeSH Terms] OR "therapeutics"[MeSH Terms] OR "treatment outcome"[MeSH Terms] OR "therapeutic use"[Subheading] OR "therapy"[Subheading]

- Test

  "diagnostic techniques and procedures" [MeSH Terms] OR "psychological tests" [MeSH Terms] OR " toxicity tests" [MeSH Terms] OR "dental caries activity test" [MeSH Terms] OR "dental pulp test" [MeSH Terms] OR "genetic complementation test" [MeSH Terms] OR "maternal serum screening tests" [MeSH Terms] OR "mutagenicity tests" [MeSH Terms] OR "radioimmunosorbent test" [MeSH Terms] OR "mandatory testing" [MeSH Terms]

We would like to point out that although the queries were manually constructed, this what not done with knowledge of the topic set. The Diagnosis and Treatment queries are so-called meta-terms (groupings of MeSH terms that correspond to a particular biomedical domain or search strategy) that were created as part of the CISMeF project[2][3]. The Test query was manually created for

[3]These and other meta-terms can be found at `http://doccismef.chu-rouen.fr/liste_des_meta_termes_anglais.html`

this competition by a medical expert. It was designed to focus on patient care. Table 1 shows the number of documents in each subcorpus after filtering. Some documents appeared in all three subsets.

Clinical question type	**# of documents**
Diagnosis	179,344
Treatment	126,026
Test	121,111

Table 1: Size of clinical question subcorpora in # of documents

2.2 Symptom extraction and disease hypotheses generation

To bridge the gap between case report and MeSH index terms, we built a Diagnostic Clinical Decision Support System (a.k.a. Symptom Checker System) to generate hypotheses of possible diseases or medical conditions for a given case report. Our system is a simple retrieval engine (Terrier using the BM25 ranking algorithm) in which diseases and their associated symptoms have been indexed. It does not contain a rule-based component or an inference engine over previous case reports. Two knowledge bases were used to construct the index, i.e. OrphaNet[4] and the Disease-Symptom Knowledge Database[5](DSKD) [3]. OrphaNet covers 6942 rare diseases; the DSKD covers 150 very frequent diseases. This is not an exhaustive list of possible diagnoses. We converted the UMLS terms in the DSKD for diseases and their symptoms to their associated MeSH terms using the "Restrict to MeSH" algorithm [1]. OrphaNet utilizes a combination of MeSH and UMLS terms as well as vocabulary from other sources. For these terms we tried to find the related MeSH terms through "Restrict to MeSH" or related entry terms in the MeSH vocabulary. We indexed both the MeSH terms and the individual words that make up the terms.

We extracted MeSH terms from the case reports (topic queries) by annotating them with MetaMap[6] and transforming the output with "Restrict to MeSH". These were then provided to the Disease-Symptom Checker both as MeSH terms and as individual words. For each case report we collected the top 5 diagnoses outputted by the system.

2.3 Age and gender extraction

We wrote a short perl script to extract gender and age information from the case reports. The script matches name variants, e.g. `"(females?|girls?|wom[ae]n)"` and outputs the relevant MeSH term, e.g. "female"[MeSH terms]. For the age groups, we only kept 5 groups: Infant, Child, Adolescent, Adult and Aged_80_and_over. For each age group the neighbouring groups were also included in the eventual query with OR operator so the final query would not be too restrictive. Overall the extraction worked very well. For three topics we were not able to extract gender information since this was not explicitly mentioned in the case summaries.

[4]Orphadata: Free access data from Orphanet. ©INSERM 1997. Available on `http://www.orphadata.org` [4]. Data version 1.0.20

[5]Freely accessible at `http://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html`

[6]Can be found at `http://metamap.nlm.nih.gov/`

2.4 Retrieval

Two different retrieval engines were used for the reported experiments.

2.4.1 Terrier

For the bag-of-words components of runs 1, 4 and 5 we used the Terrier search engine[7] (version 3.6) with an individual index for each of the three subcorpora described in Table 1. We did not apply any stemming but did carry out stopword removal in indexing and searching processes for our runs. All runs consisted of a probabilistic retrieval model based on the BM25 scoring function with the default parameters.

We indexed all fields of the original NXML files except the `<journal-meta>` and dependent fields, the `<contrib-group>` from `<article-meta>` and the `<back>` and dependent fields.

We noticed that some of the PMC identifiers were not indexed by Terrier since these codes violate the second tokeniser condition ("2. Any term which has more than 4 digits is discarded.") We therefore defined our own tokenizer which also allowed more than 3 consecutive identical characters to occur.

2.4.2 Retrieval using PubMed

The retrieval results in runs 2 and 3 were obtained by querying PubMed directly and then converting the retrieved PubMed Identifiers (pm_id) to their PubMedCentral (pmc) equivalents. Since the Open Access set had grown (compared to the 21-01-2014 snapshot provided by the organizers), we filtered out the more recent articles. The rankings of the retrieved results are taken directly from PubMed's "Sort by Relevance" feature.

3 Submitted runs

As mentioned above, all 5 runs use the same approach of selecting documents by clinical question type. In runs 2 and 3 MeSH queries are constructed which are used to query PubMed directly. Runs 1, 4 and 5 are BoW runs on the indexed subcorpora in Terrier.

Table 2 gives an overview of the components used for the different runs.

	Representation of medical case				**Strategy for Clinical Question type**	
	plain text	gender/ age	symptoms	disease hypotheses	Text in relevant subcorpus	MeSH query
Run1BoWC	text				Terrier BM25	
Run2MeSHDi		MeSH		MeSH		PubMed
Run3MeSHDiCa		MeSH	MeSH	MeSH		PubMed
Run4BoWDiCa	text			terms (name variants)	Terrier BM25	
Run5BoWDiCaS	text		terms (name variants)	terms (name variants)	Terrier BM25	

Table 2: System components for different runs

[7] Can be downloaded at `http://terrier.org/`

Run1BoWC is a baseline run. The individual words from the case report form the query that is used to query the relevant subcorpus index (diagnosis|test| treatment) in Terrier.

In **Run2MeSHDi** each case report is put through the Symptom Checker and the top 5 of the resulting hypotheses are transformed into MeSH terms. These are combined (with OR) to form a basic query. In a second step, the age group and gender are extracted from case report, transformed into their respective MeSH terms and added to the query. Finally we combine the query with the manually made MeSH query for the relevant clinical type. We use the final MeSH to query PubMed directly.

Run3MeSHDiCa is identical the procedure for Run2 except that the final MeSH queries also contain the MeSH terms for symptoms extracted from the case reports which were selected using MetaMap and the "Restrict to MeSH" algorithm.

In **Run4BoWDiCa** each case report is put through the Symptom Checker and the name variants (extracted from UMLS, OrphaNet and DSKB) from the top 5 hypotheses are combined with the words from the case report to form Bag-of-Word queries. We then performed text-based retrieval in the relevant subcorpus index (diagnosis|test| treatment) in Terrier.

Run5BoWDiCaS is identical to the procedure for Run4 except that the final final BoW queries also include the name variants for the symptoms extracted from the case reports.

4 Results

Table 3 summarizes the results of our official submitted runs as well as one additional run, according to the following official measures: bpref, R-prec and P10. The last column shows the number of relevant documents retrieved. We observe that the best among the official submitted runs is the run in which MeSH queries based on symptoms and disease hypotheses extracted from the case reports were used to query PubMed directly.

Run name	**bpref**	**R-prec**	**P10**	num_rel_ret
Run1BoWC	0.0104	0.0061	0.0100	63
Run2MeSHDi	0.0106	0.0077	0.0133	41
Run3MeSHDiCa	**0.0177**	**0.0135**	**0.0433**	**107**
Run4BoWDiCa	0.0098	0.0076	0.0167	67
Run5BoWDiCaS	0.0067	0.0039	0.0100	45
RunMeSHDiCa	0.0168	0.0122	0.0467	109

Table 3: Retrieval scores for official runs

Compared to the other participants we achieved relatively low scores: For only 6 out of 30 topics were our P10 scores equal or higher to the median of the scores of all participants for that topic.

Preliminary analysis of the official results shows that these low scores are -partly- caused by our too strict filtering approach in selecting documents per clinical question type. Table 1 shows that the three subsets comprise 236,846 documents in total which means that around 67% of the

original corpus was not included. In regards to the number of relevant documents (based on the qrels that were made available after the competition), the filtering on clinical question type had a devastating effect on recall: In the diagnosis, treatment and test subcorpus respectively only 35%, 20% and 36% of the relevant documents were present. While the idea of using these MeSH terms to find documents specific to the clinical questions still seems worthwhile, it would be better to use them for reranking purposes than for strict filtering.

To investigate the impact of the disease hypothesis generation we reran Run3 but this time did not add the MeSH terms for disease hypotheses. The scores of this run can be found in the bottom row of Table 3. We can see that the scores are very close to the best performing run (Run3MeSHDiCa) which shows that the impact of the disease hypothesis generation is minimal. A further analysis of the MeSH terms associated with the documents in the relevance assessments is needed to determine to what extent the disease generation method outputted incorrect terms and/or missed correct terms. We attribute its lack of impact -in part- to the incomplete coverage of potential diseases and conditions by the Symptom Checker.

5 Conclusion

This paper discussed LIMSI's participation in the 2014 Clinical Decision Support Track. We opted for an exploratory approach in which we tested the impact of retrieval systems based on Bag of Words versus MeSH index terms. The highest scoring official run was a MeSH run which combined MeSH terms extracted from the case reports with those of the top 5 disease hypotheses generated from the case reports. To solve the problem of relevancy in terms of clinical question type we performed filtering on selected MeSH terms which proved too strict and encumbered recall. Though our approach did not yield good results we see it as a good starting point for future participation in the track.

References

[1] Olivier Bodenreider, Stuart J Nelson, William T Hole, and H Florence Chang. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. In *Proceedings of the AMIA symposium*, page 815. American Medical Informatics Association, 1998.

[2] Aurélie Névéol, Lina F Soualmia, Magaly Douyère, Alexandrina Rogozan, Benoît Thirion, and Stéfan J Darmoni. Using CISMeF MeSH "Encapsulated" terminology and a categorization algorithm for health resources. *International journal of medical informatics*, 73(1):57–64, 2004.

[3] Xiaoyan Wang, Amy Chused, Noémie Elhadad, Carol Friedman, and Marianthi Markatou. Automated knowledge acquisition from clinical narrative reports. In *AMIA Annual Symposium Proceedings*, volume 2008, page 783. American Medical Informatics Association, 2008.

[4] Steffanie S Weinreich, R Mangon, JJ Sikkens, ME Teeuw, and MC Cornel. Orphanet: a European database for rare diseases. *Nederlands tijdschrift voor geneeskunde*, 152(9):518–519, 2008.